A REVIEW ON PREDICTION AND DIAGNOSIS OF DIABETES MELLITUS USING DATA MINING ALGORITHMS

Divya Agrahari, M.Tech Scholar, Department of Computer Science & Engineering, Buddha Institute of Technology GIDA , Gorakhpur, India.
Dr. Anshu Kumar Dwivedi, Associate Professor, Department of Computer Science & Engineering, Buddha Institute of Technology GIDA, Gorakhpur, India.

Abstract—A Diabetes, often known as diabetes mellitus (DM), is a dangerous condition that affects people all over the world. Obesity, a high blood glucose level, a lack of physical activity, and other risk factors can all lead to diabetes. If it is discovered at an early stage, there is a chance that it can be managed. The development of a computer system or program that is able to modify itself and learn from previous experiences is an example of machine learning. PIMA dataset is utilized during the course of this investigation. The collection includes around 9 characteristics for each of 768 patients. There are numerous varieties of each algorithmic approach to machine learning However, for the purposes of these study works, we select three methods that are not supervised learning. Logistic regression, decision tree, and random forest are the names of the algorithms. Every single one of these algorithms model were trained and put through their paces. In the end, we will compare and examine the performance of the using some sort of metric algorithmic approaches to machine learning. Accuracy, F-measure, Recall, and Precision are the several performance metrics that are evaluated. The Logistic Regression model has the best accuracy score, which is 74%, as well as the highest precision value, 0.73, and it also has the highest overall score hold the highest value of 0.68 for their f-measure. The best recall score was achieved by Decision Tree, which was 0.61.

***Keywords*—** Data mining, Diabetes Mellitus Disease, EM algorithm, Random Forest with Feature Selection, Machine Learning Algorithm, etc.

I. INTRODUCTION

Diabetes is an illness that can lead to death. Obesity, high blood glucose levels, a sedentary lifestyle, a lack of physical activity, and other factors can all contribute to the development of diabetes. It has an effect on the hormone insulin, which causes crabs to have an irregular metabolism and increases the amount of sugar in the blood. Diabetes results when the body fails to produce sufficient amounts of insulin. According to the World Health Organization, there are around 422 million individuals who are afflicted with diabetes. This number is disproportionately high in countries with low or no wealth. And it's possible that by the year 2030, this number will have climbed to 490 billion. On the other hand, the prevalence of diabetes can be seen in a number of different countries, such as Canada, China, and India. Diabetes is a substantial contributor to death rates all around the world. Early detection of diseases such as diabetes enables patients to receive treatment and potentially saves their lives. In order to achieve this goal, this research investigates the possibility of diabetes prediction by taking into account a variety of characteristics that are associated with the diabetes condition. To do this, we make use of the Diabetes Dataset and apply a variety of machine learning classification and ensemble techniques. The end result is an accurate prediction of diabetes.

Different strategies, including insulin and food, can be used to manage diabetes. It should be recognized for this as soon as feasible, and the proper therapy should then be given. Chemical and physical testing serves as the foundation for the majority of classification, identification, and diagnosis procedures. A specific disease can be anticipated using the inference drawn from these results. Predictions could be wrong. This is brought on by the varying degrees of uncertainty in the various testing parameters [2]. These uncertainties impede the possibility of a disease cure and lead to inaccurate predictions. A lot of development has been made in the computing facility. These developments in information technology allow for more accurate data classification, outcome prediction, and disease diagnosis in many cases. The fundamental benefit of information technology is that hospitals continuously retain and monitor large data storage of prior patients' records for numerous references [3]. The doctors can investigate various patterns in the data set with the use of these medical data. The classification, prognosis, and diagnosis of diseases may be aided by the patterns discovered in data sets [4].

The term "machine learning" refers to a training process that can be used to either computers or machines. A wide variety of machine learning techniques produce effective results for knowledge collection by constructing a wide variety of classification and ensemble models using the data that has been collected. The diabetes risk can be estimated with the help of such obtained data. There are many different approaches to machine learning, each of which is capable of prediction; nevertheless, it can be difficult to determine which approach is the best. Therefore, in order to accomplish this goal, we apply well-known classification and ensemble algorithms to the dataset in order to make a forecast.

The concept of machine learning refers to the process by which computers figure out how they can complete jobs without being specifically trained to do so [5]. It is the process of computers learning from the data that is presented in order for them to carry out specific jobs. For relatively straightforward jobs that are delegated to computers, it is possible to write algorithms that instruct the device how to carry out each step necessary to solve the issue at hand; in this case, the computer does not need to engage in any form of learning. When it comes to more complex activities, it can be difficult for a human to manually build the necessary algorithm. It is possible that assisting the machine in the development of its own algorithms will prove to be more productive in practice than having human programmers manually specify each step that is required.

The search for artificial intelligence was the impetus behind the development of machine learning as a field of study. During the early stages of artificial intelligence's development as an academic field, a number of academics showed an interest in teaching computers to learn from data. They attempted to approach the issue using a variety of symbolic methods in addition to what was at the time referred to as a "neural network." The "neural network" consisted primarily of the perceptron and other models that were later discovered to be reinventions of the generalized linear model of statistics. In addition to that, probabilistic reasoning was utilized, in particular for computerized medical diagnosis.

Diabetes is a problem that is killing a significant number of individuals all over the world. The improvement of human life is directly correlated to the development of various technologies. Therefore, there is no reason not to make advantage of the technologies to make living a healthy lifestyle easier. The technologies of deep learning and a variety of machine learning algorithms are utilized in many different kinds of prediction capabilities. Frequently employed by the world's most successful companies to increase profits and sales, the question of how humanity can benefit from the application of these technologies has been posed to us here [6]. The many algorithms that we have used and learnt over the course of our history are going to be put to the test in order to make a prediction about something that is so specialized that only specialists can do it. In order for the machine to learn the complexities of the many different aspects of the biomechanics of human beings and to accurately foresee the difficult challenges faced by live beings, it needs to be trained with the minds of medical professionals. Implementing these algorithms is necessary in order to provide accurate predictions regarding complex diseases by making use of a wide variety of internal and external characteristics that are derived from a reliable dataset [7].

Diabetes is a disorder that affects human body by lowering the insulin that delivers glucose into the blood cells. Diabetes is a lifelong chronic condition that affects human body. This causes a rise in the sugar level in the body, which can lead to a variety of health problems and even death, including stroke, heart disease, blindness, and renal failure. Patients suffering from diabetes typically exhibit the following symptoms.

- An increase of thirst experienced
- Nausea and vomiting
- Infections with a sluggish recovery time
- A greater degree of hunger
- Haze in the eyes
- Reduced body weight
- Frequent urinating

The diagnosis of diabetes mellitus relies on the following series of medical investigations.

- Urine test
- Fasting blood glucose level
- Random blood glucose level
- Oral glucose tolerance test
- Glycosylated hemoglobin(HbAlc)

II. Literature Survey

In this section, we are going to take a close look at some of the previous research that is relevant to this topic.

The research work of Jitranjan Sahool et al. [3] predicting diabetes using Machine Learning Classification Algorithms and this research work shows that, Logistic regression was found to outperform all of the machine learning algorithm showing the maximum accuracy of 72.17% in comparison to other algorithm.

Nonso et al. [4] introduced a new method for predicting the start of diabetes: in the supervised learning approach, five commonly used classifiers are utilized for the ensembles, and a meta-classifier is utilized to aggregate the outputs of the individual classifiers. The results that are presented are compared with the findings of other research that have been published that were conducted using the same dataset. It has been demonstrated that the proposed strategy can result in a greater level of accuracy when applied to diabetes onset prediction.

Tejas et al. reported Diabetes Prediction in a study that was done [2]. The purpose of using machine learning techniques is to predict diabetes using a variety of supervised machine learning methods such as support vector machine (SVM), logistic regression, and artificial neural network (ANN). Their work project proposes a method that is both effective and efficient for diabetic disease identification at an earlier stage.

Data mining was suggested as a novel strategy for diabetic illness prediction in a separate study carried out by Deeraj and colleagues [1]. An Intelligent Diabetes Disease Prediction System is being developed, which provides an analysis of diabetes disease by making use of a database consisting of diabetes patients. They suggest the usage of algorithms such as Bayesian and KNN (K-Nearest Neighbor) in this system to apply on a database of diabetes patients and evaluate them by taking numerous diabetes-related characteristics into consideration for the purpose of making a prediction regarding diabetes disease.

Comparisons were made between the several machine learning methods (support vector machine, logistic regression, decision tree, K-nearest neighbor, and random forest) that were used in the diabetes prediction study by Mitushi Soni et al. [5].

Various machine learning algorithms were utilized by our team. Which of These Models Is Best: Logistic Regression, Random Forest, or Decision Tree?

Methods for predicting the age at which type 1 diabetes will first manifest them have been the subject of additional research. Early childhood exposure to respiratory infections has been demonstrated to be associated with a greater risk of autoantibody seroconversion in children who come from families with a history of type 1 diabetes [8]. This was discovered through continuous monitoring of the subjects' islet auto antibodies during the first three years of their lives [8]. Longitudinal autoantibody measures have also been utilized as a risk predictor in families that have a first-degree relative with type 1 diabetes [37], in general populations [38], and in individuals who have been identified as being at risk [11]. In addition, genetic variables and genetic risk scores were utilized in order to determine whether or not islet auto antibodies were present in children who possessed high-risk HLA genotypes [9]. After a challenge, the levels of post-challenge C-peptide start to drop dramatically six months before diagnosis [6]. This is indicated by an examination of metabolic alterations. A composite risk score model that included clinical, genetic, and immunological variables was developed for high-risk children (who were tracked from birth until 9 years of age). This model showed a considerably improved T1D prediction in comparison to auto antibodies alone [10]. However, beyond the scope of the study described above, there is a dearth of application of machine learning approaches to the process of generating models of age at onset of type 1 diabetes. This is in spite of the fact that numerous researches have used a variety of machine learning techniques for type 2 diabetes [10]. Therefore, the work that is being proposed is different from the research that has been done in the past because it models the age at which type 1 diabetes (T1D) first appears in children. To do this, it makes use of statistical and machine learning models to determine the risk factors and build a predictive model.

III. DATASET DESCRIPTION

The Pima Indian Diabetic Set, which can be found in the Repository of Machine Learning datasets at the University of California, Irvine (UCI), served as the basis for the data set that was selected for categorization and experimental simulation. Patients under consideration are members of the Pima Indian people that are now residing in the state of Arizona, in the United States. Diabetes affects more than half of the Pima Indian population, and the condition is almost entirely caused by the people's excessive body fat. Numerous studies conducted on these populations have conclusively demonstrated that obesity is the primary factor in the development of diabetes. The data collection in question mostly consists of 9 properties, and there are 768 instances total [4]. Table 1 has a listing of these eight characteristics, along with their corresponding symbols.

In this research, the performance of three different approaches to diagnosing diabetes mellitus using PIDD was evaluated using a medical dataset that was obtained from the machine learning data repository at the University of California, Irvine. The dataset URL is archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. The data set has totally 9 attributes.

Table 1: Dataset description

Type	**Classification**	**Origin**	**Laboratory**
Features	8	(Real / Integer / Nominal)	(8 / 0 / 0)
Instances	768	Classes	2

Each algorithm requires the data to be presented in a specified format. The raw data should be converted into machine understandable format and this process is commonly called as pre-processing. The steps to be performed during pre-processing are the transformation of the attributes in the database to a single scale and replacement of all missing values in the data. The raw data can be stored in different formats including text, Excel or other database files. In most of the situations raw data is not in any standard format. Formatting the data to a format understandable by the algorithms can result in better time efficiency in respect to processing of the data. In most cases, rows represent a single case and columns denote the attributes of the case. In some databases the data are present in Comma Separated Values (CSV) format. That is, all the attributes are separated by commas and the presence of two consecutive commas stands for a missing data attribute. Sometimes, a question mark can be found for a missing attribute instead of empty space.

Figure 1 depicts the distribution of attribute values in relation to the class attribute labelled "0 or 1." The number of people who have diabetes is shown by the blue colour. It is evident from the figure that the majority of diabetic patients who are pregnant have values between 0 and 1.5, have plasma in the range of 99.5 to 103.5, have pressure in the range of 65 to 71, have skin fold thickness between 0 and 7, have insulin levels between 0 and 50, have a BMI between 27 and 30, have pedigree function between 0.25 and 0.50, and are between the ages of 21 and 25.

Table 2: Dataset Description

S. No.	**Attributes**
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI(Body Mass Index)
7	Diabetes Pedigree Function
8	Age

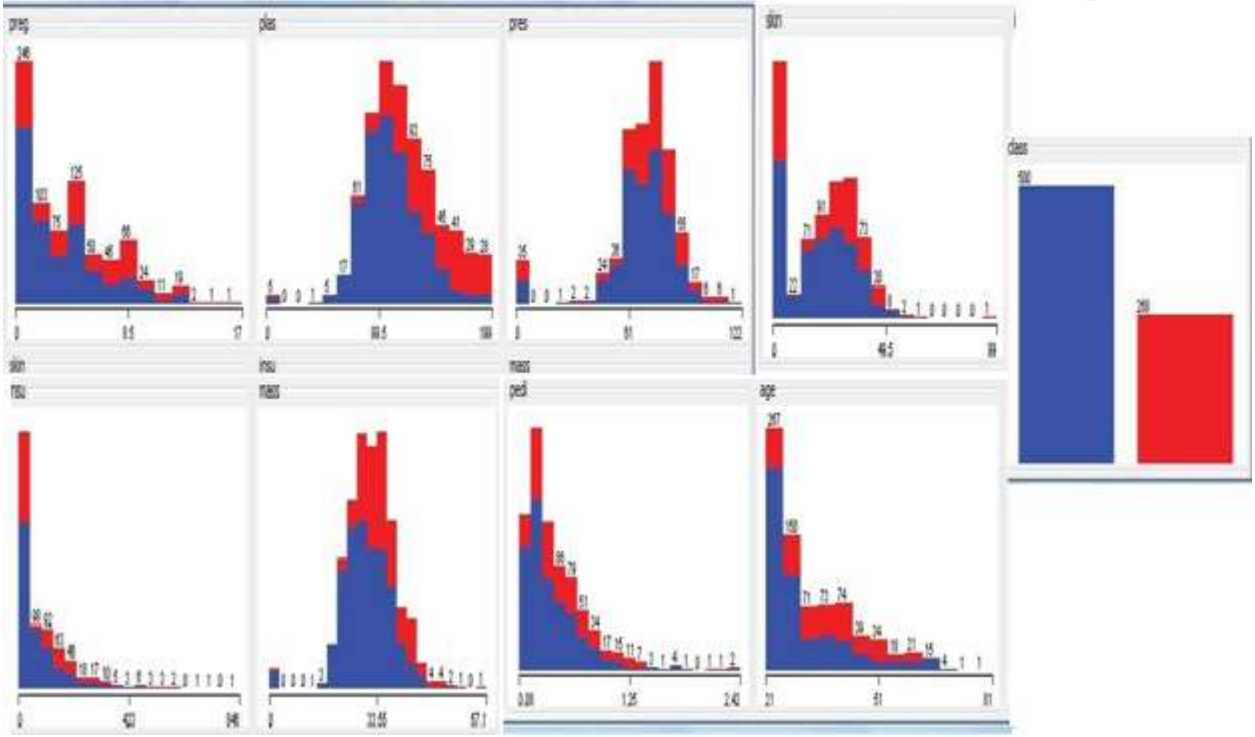


Figure 1: Attribute value distribution of Pima database

IV. DATA PREPROCESSING

The preparation of data is considered to be one of the most crucial steps. The majority of data pertaining to healthcare has missing values as well as other contaminants that can reduce the effectiveness of the data. Data preprocessing is done in order to increase the quality and effectiveness of the results obtained after the mining process.

This procedure is vital for producing reliable results and making accurate predictions in order to make optimal use of Machine Learning Techniques on the dataset. As a result, particular procedures are carried out in order to transform the data into a compact and clean data set. Before beginning the process of Iterative Analysis, you will first complete this approach. The term "Data Preprocessing" refers to the series of actions that are taken [11].

Included in it are

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

The availability of real-world data that has not been prepared makes it necessary to perform data preprocessing. The majority of real-world data is made up of erroneous information (missing data), which can occur for a variety of reasons. Some of these causes include the fact that data is not continually collected, an error in the data entry process, technological issues with biometrics, and many others.

The occurrence of noisy data (including erroneous data and outliers) - The reasons for the existence of noisy data could be due to a technological problem with the device that collects the data, a human error made when entering the data, or a number of other factors.

Data Inconsistent: The presence of inconsistencies can be attributed to a number of factors, including the existence of duplication within the data, the entry of the data by humans, the inclusion of inaccuracies in codes or names, a violation of the data constraints, and a great deal more.

In order to properly prepare our dataset on diabetes, we will need to do preprocessing in two stages.

Elimination of Missing Values: Get rid of all the occurrences where there is a value of zero (0). It is not feasible to have zero as one's worth. As a result, this particular occurrence is no longer relevant. We make a feature subset by removing irrelevant characteristics or instances, and this process is called feature subset selection. This helps to lower the dimensionality of the data, which in turn makes it easier to get things done more quickly.

Data Splitting: After the data have been cleaned, the data are split and normalized in both the training and the testing phases of the model. After the data has been split, the algorithm is trained using the training data set, and the test data set is stored separately. The training model will be generated by this procedure based on the logic and techniques used, as well as the values of the features contained in the training data. The primary purpose of normalization is to achieve the goal of bringing all of the qualities to the same scale.

V. GENERATION OF TRAINING AND TESTING DATA

The data that we work with is typically divided into two categories: training data and test data. A known output is included in the training set, and the model is trained using this data in order to be able to generalize its findings to additional data in the future. We have the test dataset, also known as the test subset, so that we may test the accuracy of our model's prediction using this subset. The size of the test set will be reduced while the training set will be increased.

The test set will be used to conduct training and testing on the training set. On the basis of certain diagnostic measurements that are contained within the dataset, the purpose of the dataset is to make a diagnostic prediction as to whether or not a patient suffers from diabetes.

VI. ALGORITHMS

MACHINE LEARNING

Machine learning is a technique of statistical learning in which every instance in a dataset is characterized by a collection of features or attributes. This technique was developed by the IBM Watson Research Center. In contrast, the term "Deep Learning" refers to a statistical learning process that extracts characteristics or qualities from raw data. This method is also known as "supervised learning."

This is accomplished using Deep Learning, which makes use of neural networks that include many hidden layers, large amounts of data, and significant processing resources. The words appear to be interchangeable to some extent; but, when utilizing the Deep Learning approach, the programme automatically creates representations of the data. In contrast, data representations in machine learning algorithms are hard-coded in the form of a set of features, which necessitates additional operations such as the selection and extraction of features (such as PCA).

Both of these phrases stand in stark contrast to an additional category of traditional artificial intelligence algorithms known as Rule-Based Systems. In these systems, every decision is manually programmed in such a way that it mimics a statistical model.

There are numerous different models that may be used for machine learning and deep learning, and these models can be divided into two distinct categories: supervised and unsupervised. Unsupervised learning makes use of methods like k-means, hierarchical clustering, and Gaussian mixture models to make an attempt to discover significant structures within the data. Learning through supervision requires assigning an output label to each individual instance contained in the dataset.

This output may have discrete or categorical values, or it may have real-valued values. The outputs of regression models are estimated to have real values, whereas the outputs of classification models are estimated to have discrete values.

Binary classification models are the simplest kind, and they only have two possible output labels: 1 (positive) and 0. (negative). The terms "linear regression," "logistic regression," "decision trees," "support vector machines," and "neural networks" are all examples of popular supervised learning algorithms that fall under the umbrella of "machine learning." Non-parametric models such as "k-nearest Neighbors" are also included in this category. Within the scope of this investigation, we will be using a supervised learning approach.

LOGISTIC REGRESSION

Logistic regression is a statistical approach that is used to analyze a dataset in which there are one or more independent variables that predict an outcome. This type of dataset might have any number of determining factors. The result is evaluated using a dichotomous variable, which means that there are only two viable interpretations of the data. Given a number of independent factors, it is used to make a prediction about whether the outcome will be binary (1/0, Yes/No, True/False). We make use of dummy variables to express outcomes that are either binary or categorical. Logical regression can also be viewed as a specific instance of linear regression that applies when the outcome variable being studied is categorical and the log of probabilities is being used as the dependent variable in the analysis. To put it another way, it calculates the likelihood that a certain event will take place by applying the collected data to a valid statistical function.

David Cox, a statistician, first invented the concept of logistic regression in the year 1958. The likelihood of a binary response can be estimated with the help of this binary logistic model, which takes into account one or more predictor factors, also known as independent variables (features). It makes it possible to state that the presence of a risk factor raises the chance of a particular result by a predetermined percentage.

Sigmoid function $P = 1/1+e - (a+bx)$

Here P = probability, a and b = parameter of Model.

RANDOM FOREST

It is a classification and regression method in addition to being a member of the ensemble learning family of approaches. This approach is quite adept at dealing with huge datasets. Leo Bremen is the creator of the Random Forest algorithm. It is a well-known method for learning in ensembles. By lowering the overall variance, Random Forest is able to improve the performance of Decision Tree. During the training phase, it builds a large number of decision trees, and then at the end of that phase, it produces the class that corresponds to the mode of the classes, also known as classification, or the mean prediction, also known as regression, of the individual trees.

1) The first step is to select the “R” features from the total features “m” where R<<M.
2) Among the “R” features, the node using the best split point.
3) Split the node into sub nodes using the best split.
4) Repeat a to c steps until ”l” number of nodes has been reached.
5) Built forest by repeating steps a to d for “a” number of times to create “n” number of trees.

The first thing that must be done is to examine the available options, then utilize the basis of each arbitrarily generated decision tree to make an educated guess as to the outcome, and last, save that educated guess in various locations throughout the desired location. The second step is to tally up the votes cast for each of the forecasted goals, and the third and last step is to use the forecast with the most votes as the basis for the ultimate prediction produced by the random forest method. Some of Random Forest's options, which can produce accurate predictions for a variety of different applications, are available to the user.

DECISION TREE

A decision tree is a specific kind of strategy for classifying data. It is a method of learning through supervision. When the response variable is categorical, the decision tree is utilized. The classification process is described by a decision tree, which is based on a model that is structured like a tree and defines input features. Variables that are input can be of any sort, including graphs, texts, discrete values, continuous values, and so forth.

1) Build the tree with the nodes serving as the input feature.
2) Select the feature to forecast the output from the input feature that has the largest information gain.
3) The information gain that is considered to be the greatest is determined for each characteristic and each node of the tree.
4) Repeat step 2 in order to build a sub tree utilizing the feature that was not used in the node that was previously examined.

K-MEANS ALGORITHM

Unsupervised algorithms are algorithms that can function on unlabeled samples without direct supervision. This indicates that the output cannot be predicted even if the input can be identified. The K means algorithm is one of several that fall under the category of unsupervised learning algorithms. They require an input parameter, the number of clusters, as well as n objects in the data collection, which is then partitioned into k clusters. A random selection of k items is made by the algorithm. Each object is given a place in one of the clusters that it belongs to according to how closely it is located to its related cluster. The following step is to locate the locations that are the most adjacent to one another. It is recommended to use the Euclidean distance while trying to locate the object's most central location. After the items have been divided up into k clusters, the new centers of the clusters are determined by taking the average of the objects within each of the k clusters in turn. This procedure is carried out until there is no longer any variation in the k cluster centers. The sum of squared error (SSE) is the objective function that the K-means algorithm seeks to minimize in order to achieve its goal [14]. The acronym SSE stands for

$$\text{argmin}_C \quad \sum_{i=1}^{k} \sum_{p \in Ci} |p\text{-}m_i|^2 \qquad (1)$$

Here, E is the total of the square errors of all of the objects that have cluster means for each of the k clusters, and p is the object that belongs to one of the clusters. The combination of *Ci* and *mi* represents the cluster mean. The total number of records in the dataset is denoted by "n," and "k" indicates the number of clusters.

Input: D is input -data set.
Output: Output is k clusters.
Step 1: Initialize cluster centers as D.
Step 2: Randomly choose k objects from D.
Step 3: Repeat the following steps until no change in cluster means/ min error E is reached.
Step 4: Consider each of the k clusters. Compare the mean value of the objects in the clusters for initialization.
Step 5: Initialize the object with most similar value from D to one of k clusters.
Step 6: Take the mean value of the objects for each of k cluster.
Step 7: Update the cluster means with respect to object value.

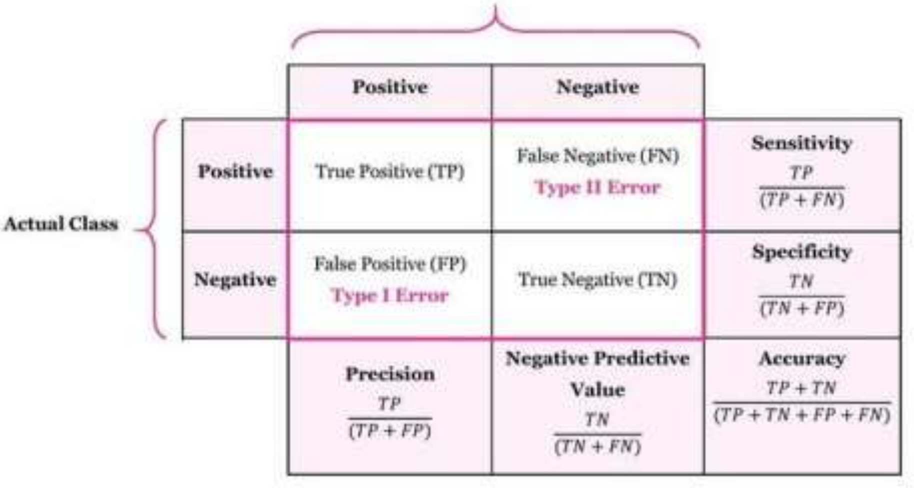
VII. PERFORMANCE MEASUREMENT

We will test our model on the dataset that we have generated, and we will also measure how well the algorithms work on the dataset that we have prepared [15]. We utilize Accuracy as a measure of the effectiveness of classifiers in order to evaluate the performance of our newly constructed classification and make it comparable to other approaches that are currently in use.
True Positives (TP) - These are the positively predicted values that turned out to be correct, which indicates that both the value of the actual class and the value of the predicted class are true. For example, if the actual class value indicates that the passenger survived and the predicted class tells you the same thing, you can assume that the passenger did indeed survive.
True Negatives (TN) - These are the negative values that have been accurately predicted, which indicates that the value of the real class is no and that the value of the predicted class is also no. For example, if the actual class reports that the

passenger did not make it out alive while the projected class reports the same thing. These results, known as false positives and false negatives, are produced when your actual class is in conflict with the class that was predicted.

False Positives (FP) – When the class that was actually taken was not the class that was projected to be taken. For example, if the actual class indicates that the passenger did not survive, but the forecast class indicates that they will, this passenger will survive.

False Negatives (FN) – The situation in which the actual class is yes while the projected class is no. For example, if the actual class value indicates that the passenger survived while the predicted class implies that the passenger would die, the actual class value should be used. When you have a firm grasp on these four characteristics, we will be able to proceed with the calculation of accuracy, precision, recall, and F-measure.



		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 2: Confusion Matrix

VIII. CONCLUSION

Concerns have been raised by medical professionals over how to identify diabetes in its early stages. In the course of this research project, an attempt was made to create a system that might predict diabetes by employing a number of different algorithms and analyzing how well they performed. The study consisted of implementing three different machine learning algorithms, and their performance was evaluated based on a variety of criteria. The PIMA Indian Diabetes dataset was used for the experiment, and the findings showed that logistic regression had the greatest performance overall. This machine learning technique is also adaptable and can be used to forecast various diseases outside the one it was designed for. The findings could be further improved by integrating other machine learning algorithms, which would improve its ability to predict diabetes.

References

[1] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[2] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13.

[3] Jitranjan Sahoo, Manoranjan Dash & Abhilash Pati, "Diabetes Prediction Using Machine Learning Classification Algorithms", International Research Journal of Engineering and Technology, Vol. 7, Issue 8, August 2020.

[4] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[5] Mitushi Soni, 'Diabetes Prediction using Machine Learning Techniques', International Journal of Engineering Research & Technology, Vol. 9, Issue 9, September 2020..

[6] Sapna. S,Tamilarasi. A and Pravin Kumar.M, 2012 "Implementation of genetic algorithm in predicting diabetes", IJCSI, International Journal of Computer Science Issues, Vol. 9, Issue 2, No 4, pp. 393-398

[7] Nirmala Devi M.,Appavu alias Balamurugan S.,Swathi U.V., 2013.",An amalgam KNN to predict Diabetes Mellitus", IEEE International Conference on Emerging Trends in Computing ,Communication and Nanotechnology(ICECCN),pp 691-695

[8] Asha Gowda Karegowda and Jayaram. A. M., 2009"Cascading GA & CFS for feature subset selection in medical data mining", IEEE International Advance Computing Conference, Patiyala, India

[9] Krzysztof J.Cios, G.William Moore (2002) 'Uniqueness of Medical Data Mining', Artificial Intelligence in Medicine Journal pp 1-19.

[10] Asha Gowda Karegowda , A.S. Manjunath , M.A. Jayaram (2011) "Application Of Genetic Algorithm Optimized Neural Network Connection Weights For Medical Diagnosis Of Pima Indians Diabetes", International Journal on Soft Computing (IJSC), Vol.2, No.2.

[11] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996) 'From Data Mining to Knowledge Discovery in Databases' AAAI Vol.17, No.3 pp 37-54.

[12] Zoran Bosnic, Petar Vracar, Milos D. Radovic, Goran Devedzic, Nenad D. Filipovic and Igor Kononenko(2012) 'Mining Data From Hemodynamic Simulations for generating Prediction and Explanation Models' IEEE Vol. 16, No. 2,pp 248-254.

[13] B.M Patil, R.C Joshi, Durga Tosniwal(2010)Hybrid Prediction model for Type-2 Diabetic Patients, Expert System with Applications, 37, 8102-8108.

[14] Polat, K., Gunes, S., & Aslan, A., (2008) A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Systems with Applications, 34(1), 214–221.

[15] D. Menon, K. Schwab, D.W. Wright, A.I. Maas, and the Demographics and Clinical Assessment Working Group of the International and Interagency Initiative toward Common Data Elements for Research on Traumatic Brain Injury and Psychological Health, Position statement: definition of traumatic brain injury, Arch. Phys. Med. Rehabil., vol. 91, pp. 1637– 40, Nov 2010.